An analysis of Parkinson's Disease Handwriting databases : PaHaW and NewHandPD

Paul Lerner[a,*], Laurence Likforman[a]

[a]*Télécom ParisTech. 46 Rue Barrault, 75013 Paris, France*

ARTICLE INFO

Keywords:
Parkinson's Disease
Parkinson Handwriting Database
PaHaW
NewHandPD
Drotar
Pereira

ABSTRACT

This template helps you to create a properly formatted LaTeX manuscript.
`\beginabstract ... \endabstract` and `\begin{keyword} ... \end{keyword}` which contain the abstract and keywords respectively.
Each keyword shall be separated by a `\sep` command.

This template helps you to create a properly formatted LaTeX manuscript.
`\beginabstract ... \endabstract` and `\begin{keyword} ... \end{keyword}` which contain the abstract and keywords respectively.
Each keyword shall be separated by a `\sep` command.

foo

1. Introduction

1.1. Handwriting analysis for PD diagnosis

Parkinson's Disease (PD) is the second most common neurodegenerative disease [20]. Its diagnosis is considered to be difficult as there is a large number of symptoms which are shared among other diseases (see, e.g. [5]). However, the diagnosis is usually assessed by a neurologist after a physical examination as SPECT and CT scans are costly, invasive, and usually effective when the disease has already progressed to a mature stage [8]. Moreover, they do not distinguish PD from multiple system atrophy or corticobasal degeneration [4].

In the description of the disease made by James Parkinson in 1817, writing deficits precede walking deficits. Since then, handwriting has been used to assess PD through exams like Archimedean spiral or simple subjective rating like in the Unified Parkinson's Disease Rating Scale (UPDRS, [11]). Symptoms such as *tremor* or *micrographia* would be visible through a traditional paper-and-pencil examination. However, for a few years, researchers have been arguing that PD *dysgraphia* is larger than *micrographia* and that it's observable through *kinematic* analysis : [7] report that writing velocity and smoothness/fluency abnormalities are more frequent (above 75%) than diminished letter size (between 30%-50%). Moreover, the use of smart pens and digital tablets has allowed for a large number of feature extraction, thus statistical analysis and machine learning.

The goal for the researchers is to provide for a Clinical Decision Support System which would confirm or question the neurologist' diagnosis based on a cheap and non-invasive handwriting exam. Two major databases have emerged from this research : *PaHaW* and *NewHandPD*, see [3] and [13] (which both have been published in this journal), respectively. I call them major because there are not the only PD handwriting database (see, e.g. [19]) but several different researchers have worked on both of them (as they are publicly available). See [8, 3, 6, 10] for works on *PaHaW*, [13, 1] for works on *NewHandPD* and [12, 18, 14] for works on *HandPD*, a former version of *NewHandPD*. We won't analyze *HandPD* because it's very unclear : on the database webpage[1] and on the above cited papers it's stated that there's 18 controls and 72 PDs, although, after downloading the database we can see that the PDs are only numbered from 1 to 38 (and there's no #4). Moreover some PDs have taken up to 8, 12 or 16 spiral exams instead of 4 like the controls. Therefore I think the original database might not be available anymore or that the subjects' identifier has been corrupted.

We argue that the *NewHandPD* database is biased because of too young controls and that one should use *PaHaW*, preferably.

1.2. PD and aging

Aging is the biggest risk factor for developing PD [16]. Moreover, several symptoms of PD are shared with normal aging, such as :

- loss of dopaminergic neurons [17]. It even seems to correlate with **motor performance** as much in both controls and PDs [15].
- **bradykinesia** (i.e. slowness of movement) [9]
- Reduced levels of tibialis anterior muscle activity [2]

However, both *NewHandPD* and *PaHaW* have older PDs than controls :

- NewHandPD controls are in average 13.8 years younger than PDs.
- PaHaW controls are in average 6.9 years younger than PDs.

The code will be made publicly available here : **ADD GITHUB**

*Corresponding author

paul.lerner@telecom-paristech.fr (P. Lerner); 1.paul.lerner@gmail.com (P. Lerner)

ORCID(s): 0000-0002-0882-8684 (P. Lerner); 0000-0001-9096-7239 (L. Likforman)

[1] http://wwwp.fc.unesp.br/ papa/pub/databases/Handpd/

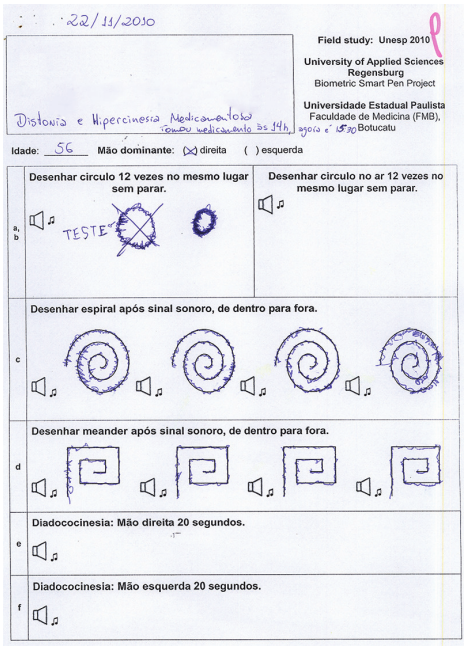

Figure 1: From [13] : Exam form for the *NewHandPD* database

We will try to study here if these gaps are significative.

1.3. Databases description

See also Appendix for some technical details about the databases.

1.3.1. NewHandPD

The well-named *NewHandPD* database is a new version of the *HandPD* database. It comprises 66 subjects (31 PDs and 35 controls). Each subject was asked to draw 12 exams, being 4 spirals (labeled *sigSp 1, 2, 3, 4* afterwards), 4 meanders (labeled *sigMea 1, 2, 3, 4* afterwards), 2 circled movements (one circle on paper, *circA* and another in the air, *circB*), and left and right-handed diadochokinesis : *sigDiaA* and *sigDiaB*, respectively (see Figure 1). See [13] for more details.

1.3.2. PaHaW

The *PaHaW* database comprises 75 subjects (37 PDs and 38 controls). Although we excluded the subjects # 46 (control), 60 (PD) and 66 (control) from this study because they didn't perform the spiral exam. Therefore in this study we present a perfectly balanced PaHaW of 36 controls and 36 PDs (as did [8, 6]). The template consists of seven handwriting tasks (see Figure 2). From the first to the third task, participants wrote cursive letters or bi/tri-grams of letters. The next three tasks involved words in Czech (the native language of participants) with the following translation to English: lektorka - teacher (female), porovnat - to compare, nepopadnout - to not catch. The final task involved a longer sentence : (Tramvaj dnes uz nepojede - The tram won't go today). See [3] for more details.

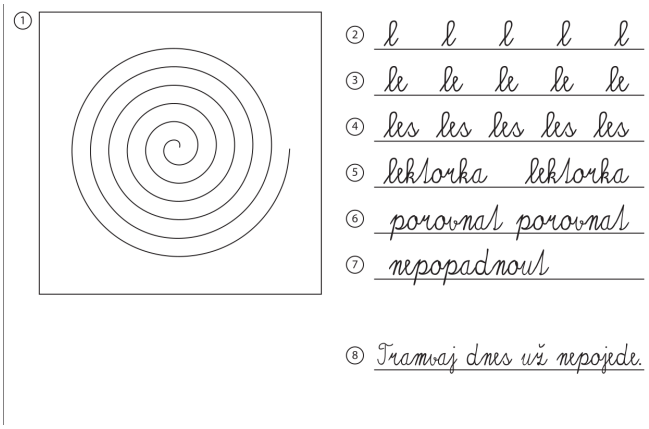

Figure 2: From [3] : Exam form for the *PaHaW* database

2. Data analysis and classification

We will first investigate a very interesting feature : the exam duration. PDs suffer from *bradykinesia* (i.e. slowness of movement) which causes them to complete a graphomotor task in more time than usually required [8]. However, you'll see that this feature is unbelievably discriminative on *NewHandPD*. First clue of this biased database. We will then explain why it is so by analyzing the subject's age on both databases.

2.1. Exam Duration

2.1.1. Statistical analysis

The *NewHandPD* database has a serious defect : the PDs exams take significantly much time than Controls' exams : **11.5 seconds** in average ! Thus when plotting the average duration of the exams per subject we can clearly see a threshold that separates almost all controls and all PDs (see Figure 3). We'll see that the phenomenon is visible on every single exam. As you can see on Figure 3, the phenomenon is not similar on *PaHaW* where only 3 PDs are clearly above the rest of the subjects.

Therefore, when using a simple rule :

- subject is PD if exam_duration > 15242 ms *(median duration over all subjects)*
- else control

We're able to achieve 85% accuracy on the *NewHandPD* database ! See section 2.1.2 for proper classification tests (i.e. with train-test sets etc.). See section 2.2 for interpretation of this defect.

I used Student's t-test to study the resemblance between the PD's exams durations and the controls' exams duration.

Before applying t-test on a database, it needs 2 properties :

1. all input samples are from populations with equal variances.
2. data was drawn from a normal distribution.

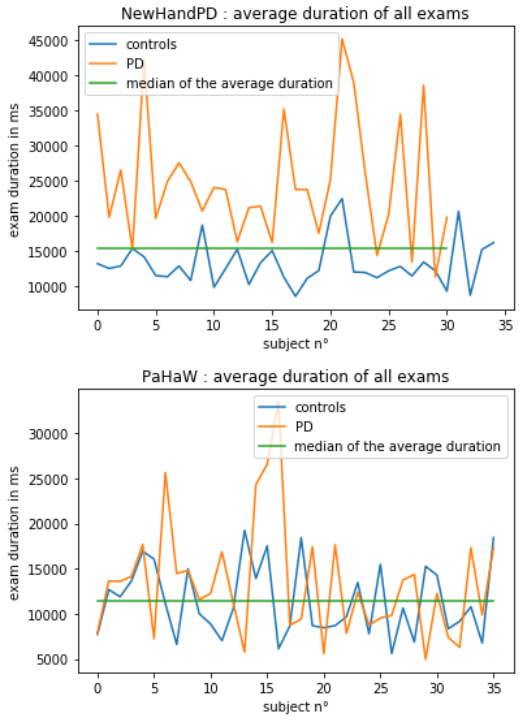

Figure 3: average duration of the exams. Up : *NewHandPD*, down : *PaHaW*.

So I applied Bartlett's test for the first condition and Shapiro-Wilk for the second using SciPy[2]. See Table 1 As you can see from the p-values the tests are strongly significant. I

Table 1
NewHandPD : resemblance between the PD's exams durations and the controls' exams duration.

exam	t-statistic	p-value
circA	6.1612	5.361e-08
circB	5.9077	1.457e-07
sigDiaB	3.8808	2.487e-04
sigMea3	6.0783	7.441e-08
sigSp1	5.2307	1.987e-06
meanders	6.4848	1.478e-08
spirals	6.2149	4.332e-08
all	7.2950	5.638e-10

meanders : average duration of the four meanders, idem for *spirals*. *all* : average duration of all exams.

didn't include sigDiaA because it failed Bartlett's test (i.e. the duration of PD's and control had different variances). Nor SigMea 1, 2, 4 nor SigSp 2, 3 and 4 because either the control or the PD group failed the Shapiro-Wilk's test (i.e. the duration was not drawn from a normal distribution.).

In order to have a comparison point I studied the same thing with the *PaHaW* database. As you can see in table 2 from the p-values, none of the tests were significant, although I excluded the *l* exam because it failed the Bartlett's test. Moreover, you'll see in section 2.1.2 that it doesn't translate into good classification accuracies.

Table 2
PaHaW : resemblance between the PD's exams durations and the controls' exams duration

exam	t-statistic	p-value
spiral	1.5826	0.1180
le	1.3512	0.1810
les	1.4268	0.1581
lektorka	1.4210	0.1598
porovnat	0.9418	0.3495
nepopadnout	1.4159	0.1612
tram	1.0952	0.2772
all	1.5799	0.1186

all : average duration of all exams.

2.1.2. *Classification*

To further confirm those results, I used scikit-learn[3] implementation of Linear Discriminant Analysis (LDA). LDA models the distribution of the features X separately in each of the response classes (i.e. given Y, here PD or control), and then use Bayes' theorem to flip these around into estimates for $\mathbb{P}(Y = k|X = x)$. I feed the model only the duration of the exam (one feature).

See table 3 for the results. In order to have comparable results with [1] I used random 50-50 train-test split with 15 runs. The evaluation method of [13] is unknown.

Table 3
NewHandPD : Accuracy (%) depending on the exam (average over the 15 runs ± std).

exam	LDA (my model)	[1]	[13]
circA	80.61 ± (4.40)	76.17 ± 6.92	68.04 ± 2.96
circB	83.43 ± (4.68)	76.69 ± 5.38	73.41 ± 3.66
sigDiaA	75.96 ± (5.90)	68.69 ± 7.26	73.59 ± 3.57
sigDiaB	77.17 ± (5.52)	66.30 ± 7.38	76.32 ± 5.18
meanders	81.82 ± (3.13)	81.07 ± 2.60	80.75 ± 2.08
spirals	82.02 ± (3.02)	81.03 ± 2.40	78.26 ± 1.97
all	83.84 ± (3.26)	NA	95.74 ± 1.60

NA : Not Applicable

I averaged the duration of the four meanders and the four spirals to provide the results of the lines *meanders* and *spirals*. Although I achieved similar results when using only one meander or one spiral to train the model. In the same way in *all* I averaged the duration of all exams. Majority voting provided similar results. We can see that it doesn't give very better results, this is because the same subjects are misclassified over the exams since we use only one feature that is very dependent on the subject.

I displayed here the results of [1] when using first OPF then SVM because it's their best results. In the same way the results of [13] are the one where they used the "CNN-ImageNet" model on "images" of size 128x128 because it's

[2] https://www.scipy.org/

[3] https://scikit-learn.org/

their best results.

Notice how we outperform both of them on every single exam although [13] achieved better results on all exams (they use majority voting).

I think we can conclude from that that the transformation of the data from sensor to image of [13] and the Discrete Wavelet Transform of [1] have probably loss the temporal information of the data.

I used the exact same model and feature (i.e. LDA and exam duration) on the PaHaW database. As I did before, in order to have a comparison point, I displayed here the results of [3] when using a SVM as it's their best results. Their result is the average over 10 fold cross-validation but they don't provide for standard deviation (std). I advise you that, in order to get comparable results with the *NewHandPD* database, I used the same evaluation method as before (i.e. random 50-50 split with 15 runs) and not 10 fold cross-validation like [3].

You can see in table 4 that [3] outperform my model by a large margin on all exams as my model barely gets above chance level on some exams and falls behind on others.

Table 4
PaHaW : Accuracy (%) depending on the exam (average over the 15 runs ± std).

exam	LDA (my model)	[3]
spiral	50.56 ± (6.97)	62.8
l	48.52 ± (7.23)	72.3
le	58.33 ± (6.25)	71
les	55.93 ± (3.91)	66.4
lektorka	48.70 ± (6.24)	65.2
porovnat	49.63 ± (6.16)	73.3
nepopadnout	50.19 ± (6.99)	67.6
tram	51.30 ± (5.73)	76.5
all	56.48 ± (4.61)	81.3

2.2. Age

As mentioned in section 1.2, not only PD is correlated with age but several symptoms are shared among PD and normal aging, including bradykinesia which we have observed in the previous section thanks to the exam duration !

Therefore, in this section we will try to explain why classification of *NewHandPD* subjects was so easy and thus why the results displayed in section 2.1.2 are biased.

2.2.1. Statistical analysis

When plotting the exam duration against the subject's age (figure 4), we can see two nice clusters of PDs and controls on the *NewHandPD* database. However on *PaHaW*, there's only a 4 PD's cluster, which explains the differences in the classification accuracies of section 2.1.2. As you can see on figure 4, on both databases, the average duration of the exam is correlated to the subject's age. That's not a discovery, see, e.g. [9]. The problem is that in the *NewHandPD* database, most controls are younger than most PDs. On the contrary, you can see that on the *PaHaW* database, no control is younger than the youngest PD and no PD is older than the oldest control.

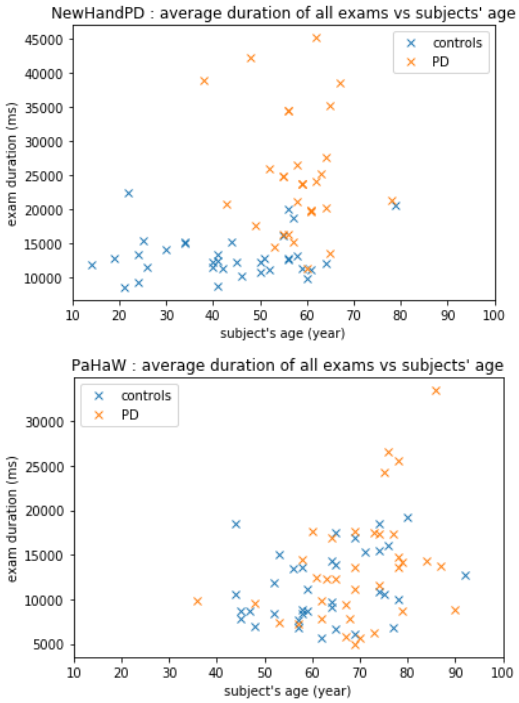

Figure 4: Duration of all exams vs. subject's age. Up : *NewHandPD*, down : *PaHaW*.

I used Pearson's correlation coefficient to measure the correlation between the exams duration and the age of the subject. Pearson's correlation coefficient measures the linear correlation between two variables. It has a value between +1 and − 1, where 1 is total positive linear correlation, 0 is no linear correlation, and − 1 is total negative linear correlation.

I used its SciPy implementation. As you can see in table 5, on *NewHandPD* the durations of all exams are positively correlated to the subject's age. The correlations of sigDia A and B are weaker than the other exams. This explains why the classifier obtain worse results on these exams (see table 3).

I performed the same analysis with *PaHaW*, see table 6. As we could expect from the plot and the literature, all exams duration are significantly correlated with subject's age.

However if we focus only on the age (plotted in figure 5 on two separate lines for better readability) we can still see some clusters on the *PaHaW* database, e.g. there's a group of 4 PDs who have between 80 and 90 years old. This translates into a significant difference between controls' and PD's age when applying a t-test and 64% accuracy when training a LDA using only the subject's age (see table 7 and section 2.2.2).

2.2.2. Classification

Moreover, as I attempted to classify the subjects using only the duration of their exam with a LDA (see section 2.1.2), here I used the same model using only the subject's

Table 5
NewHandPD : Correlation between exams duration and subject's age.

exam	pearson corr. coeff.	p-value	significance
circA	0.2781	0.0238	**
circB	0.2533	0.0401	**
sigDiaA	0.2099	0.0907	*
sigDiaB	0.2073	0.0948	*
sigMea1	0.3822	0.0015	***
sigMea2	0.3728	0.0021	***
sigMea3	0.3265	0.0075	***
sigMea4	0.3226	0.0082	***
sigSp1	0.3079	0.0119	**
sigSp2	0.3242	0.0079	***
sigSp3	0.3169	0.0095	***
sigSp4	0.3292	0.0070	***
meanders	0.3619	0.0028	***
spirals	0.3346	0.0060	***
all	0.3512	0.0038	***

significance : weak *, strong **, very strong ***

Table 6
PaHaW : Correlation between exams duration and subject's age.

exam	Spearman corr. coeff.	p-value	significance
spiral	0.2163	0.0681	*
l	0.2557	0.0301	**
le	0.3088	0.0083	***
les	0.4225	0.0002	***
lektorka	0.3086	0.0084	***
porovnat	0.2787	0.0178	**
nepopadnout	0.2575	0.0290	**
tram	0.2814	0.0166	**
all	0.3631	0.0017	***

significance : weak *, strong **, very strong ***

Table 7
Statistical Analysis of the PDs and controls' age on both databases.

database	t-statistic	p-value	significance
NewHandPD	4.85	8.0E-06	***
PaHaW	2.57	1.3E-02	**

significance : weak *, strong **, very strong ***

age. The classification evaluation is the same as before (i.e. random 50-50 split with 15 runs) therefore the results in table 8 are print as percentage ± std over the 15 runs.

3. Conclusion

These two databases have older PDs than controls. Therefore, on both databases we're able to classify subjects based on their age with an accuracy significantly higher than chance level. Moreover, Student's t-test has been applied on both databases and shows that there's a significant difference between the controls and the PDs' age. However :

Figure 5: Subject's age. Up : *NewHandPD*, down : *PaHaW*.

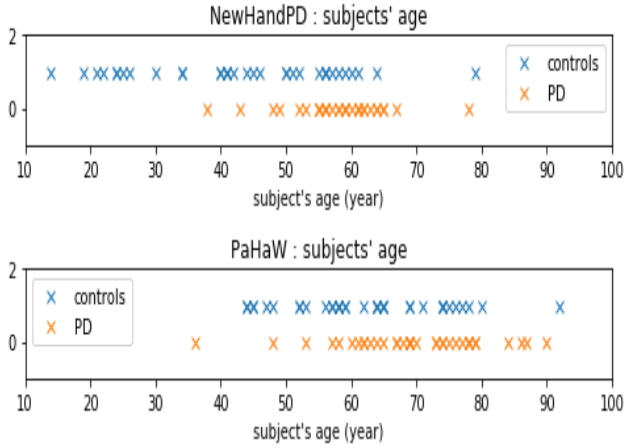

Table 8
Classification based only on subject's age on both databases.

database	LDA
NewHandPD	75.76 ± 3.67
PaHaW	63.70 ± 7.49

1. We're able to classify subjects of NewHandPD way more accurately than those of PaHaW using the same model and feature (i.e. LDA and subject's age).
2. The difference between controls and PDs' age is more significant on NewHandPD than in PaHaW.

Moreover, the duration of the exam is significantly correlated to the subject's age on both databases (that's expected, see sections 1.2 and 2.2) and :

1. We're able to classify the subjects of NewHandPD more accurately than previous researchers using only the exam' duration although with the same model and feature we barely get above chance level on PaHaW.
2. The exam duration of PDs and controls are significantly different in the NewHandPD database but are similar in PaHaW.

Therefore we advise the authors of NewHandPD to discard the younger controls and add some older controls to the database. Adding new subjects with a better age match would also be beneficial for the PaHaW database.

We also advise the readers that the results obtained on NewHandPD are biased and those obtained on PaHaW might be biased.

Acknowledgments

who funded my internship ? should we put it here ?

References

[1] Afonso, L.C.S., Pereira, C.R., Weber, S.A.T., Hook, C., Papa, J.P., 2017. Parkinson's disease identification through deep optimum-path

forest clustering, in: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE. pp. 163–169.
[2] Baker, K.K., Ramig, L.O., Luschei, E.S., Smith, M.E., 1998. Thyroarytenoid muscle activity associated with hypophonia in parkinson disease and aging. Neurology 51, 1592–1598.
[3] Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., Faundez-Zanuy, M., 2016. Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. Artificial intelligence in Medicine 67, 39–46.
[4] de la Fuente-Fernández, R., 2012. Role of datscan and clinical diagnosis in parkinson disease. Neurology 78, 696–701.
[5] Hughes, A.J., Daniel, S.E., BenâĂŘShlomo, Y., Lees, A.J., 2002. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. Brain 125, 861–870. URL: https://doi.org/10.1093/brain/awf080, doi:10.1093/brain/awf080, arXiv:http://oup.prod.sis.lan/brain/article-pdf/125/4/861/17864673/1250861.pdf.
[6] Impedovo, D., Pirlo, G., Vessio, G., 2018. Dynamic handwriting analysis for supporting earlier parkinsonâĂŹs disease diagnosis. Information 9, 247.
[7] Letanneux, A., Danna, J., Velay, J.L., Viallet, F., Pinto, S., 2014. From micrographia to parkinson's disease dysgraphia. Movement Disorders 29, 1467–1475.
[8] Moetesum, M., Siddiqi, I., Vincent, N., Cloppet, F., 2019. Assessing visual attributes of handwriting for prediction of neurological disordersâĂŤa case study on parkinsonâĂŹs disease. Pattern Recognition Letters 121, 19–27.
[9] MORTIMER, J.A., 1988. Human motor behavior and aging. Annals of the New York Academy of Sciences 515, 54–66. URL: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1988.tb32966.x, doi:10.1111/j.1749-6632.1988.tb32966.x, arXiv:https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1988.tb32966.x.
[10] Mucha, J., Mekyska, J., Faundez-Zanuy, M., Lopez-De-Ipina, K., Zvoncak, V., Galaz, Z., Kiska, T., Smekal, Z., Brabenec, L., Rektorova, I., 2018. Advanced parkinson's disease dysgraphia analysis based on fractional derivatives of online handwriting, in: 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE. pp. 1–6.
[11] on Rating Scales for Parkinson's Disease, M.D.S.T.F., 2003. The unified parkinson's disease rating scale (updrs): status and recommendations. Movement Disorders 18, 738–750.
[12] Passos, L.A., Pereira, C.R., Rezende, E.R., Carvalho, T.J., Weber, S.A., Hook, C., Papa, J.P., 2018. Parkinson disease identification using residual networks and optimum-path forest, in: 2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI), IEEE. pp. 000325–000330.
[13] Pereira, C.R., Pereira, D.R., Rosa, G.H., Albuquerque, V.H., Weber, S.A., Hook, C., Papa, J.P., 2018. Handwritten dynamics assessment through convolutional neural networks: An application to parkinson's disease identification. Artificial intelligence in medicine 87, 67–77.
[14] Pereira, C.R., Pereira, D.R., da Silva, F.A., Hook, C., Weber, S.A., Pereira, L.A., Papa, J.P., 2015. A step towards the automated diagnosis of parkinson's disease: Analyzing handwriting movements, in: 2015 IEEE 28th international symposium on computer-based medical systems, IEEE. pp. 171–176.
[15] Pujol, J., Junqué, C., Vendrell, P., Grau, J.M., Capdevila, A., 1992. Reduction of the substantia nigra width and motor decline in aging and parkinson's disease. Archives of neurology 49, 1119–1122.
[16] Reeve, A., Simcox, E., Turnbull, D., 2014. Ageing and parkinson's disease: why is advancing age the biggest risk factor? Ageing research reviews 14, 19–30.
[17] Rodriguez, M., Rodriguez-Sabate, C., Morales, I., Sanchez, A., Sabate, M., 2015. Parkinson's disease as a result of aging. Aging cell 14, 293–308.
[18] de Souza, J.W., Alves, S.S., Rebouças, E.d.S., Almeida, J.S., Rebouças Filho, P.P., 2018. A new approach to diagnose parkinsonâĂŹs disease using a structural cooccurrence matrix for a similarity analysis. Computational intelligence and neuroscience 2018.
[19] Taleb, C., Khachab, M., Mokbel, C., Likforman-Sulem, L., 2017. Feature selection for an improved parkinson's disease identification based on handwriting, in: 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), IEEE. pp. 52–56.
[20] Tanner, C.M., Ottman, R., Goldman, S.M., Ellenberg, J., Chan, P., Mayeux, R., Langston, J.W., 1999. Parkinson disease in twins: an etiologic study. Jama 281, 341–346.

A. Technicalities about the databases

A.1. On NewHandPD

The subject H34's age was set at 40 on all exams except for the SigMea1 where it was set at 50. Therefore, I assumed it was a mistake and set it at 40.

It's actually unclear which database did [13] use : at some point they say that their database comprises only 34 subjects instead of 66 but at the same time they link of the database webpage[4] where it says that the database is well containing 66 subjects (and one can see for itself after downloading the data). Moreover, [1], which is almost the same group of authors as [13], confirms that NewHandPD is comprises 66 subjects, although it was published before [13]. Therefore I think we can assume that [13] use the same database as [1] and us and that the 34 subjects they talk about were added to the *HandPD* database to form *NewHandPD*.

A.2. On PaHaW

All the subject's numbers I mention are counting from zero and after excluding the subjects that didn't perform the spiral exam.

The exams spiral, l, lektorka and tram from subjects 56, 9, 39 and 67, respectively, started their exam while pen was in air (i.e. not on paper). Although [3] say that the recording starts when the pen first touches the paper so I assume it's a recording error and discarded the in-air points.

In the same way the subject # 43 has a recording problem on his tram exam : the *timestamp* measure jumps to 10^{12} on the 12 last points of the recording although the timestamps from all subjects are in the 10^6 magnitude. So I assume it's a recording error and discarded the 12 last points.

[4] http://wwwp.fc.unesp.br/ papa/pub/databases/Handpd/